Adaptive Direct RGB-D Registration and Mapping for Large Motions

Renato Martins, Eduardo Fernandez-Moral, Patrick Rives

Adaptive Direct RGB-D Registration and Mapping for Large Motions

Renato Martins, Eduardo Fernandez-Moral and Patrick Rives

Inria Sophia Antipolis

Abstract. Dense direct RGB-D registration methods are widely used in tasks ranging from localization and tracking to 3D scene reconstruction. This work addresses a peculiar aspect which drastically limits the applicability of direct registration, namely the weakness of the convergence domain. First, we propose an activation function based on the conditioning of the RGB and ICP point-to-plane error terms. This function strengthens the geometric error influence in the first coarse iterations, while the intensity data term dominates in the finer increments. The information gathered from the geometric and photometric cost functions is not only considered for improving the system observability, but for exploiting the different convergence properties and convexity of each data term. Next, we develop a set of strategies as a flexible regularization and a pixel saliency selection to further improve the quality and robustness of this approach.

The methodology is formulated for a generic warping model and results are given using perspective and spherical sensor models. Finally, our method is validated in different RGB-D spherical datasets, including both indoor and outdoor real sequences and using the KITTI VO/SLAM benchmark dataset. We show that the different proposed techniques (weighted activation function, regularization, saliency pixel selection), lead to faster convergence and larger convergence domains, which are the main limitations to the use of direct methods.

1 Introduction

Feature based registration methods have bigger convergence domains (if feature matching is successful) but are locally less precise and more sensitive to outliers than direct dense methods [1] [2]. Feature-based methods (e.g. [3,4,5,6]) rely on an intermediary estimation process based on thresholding [7,8] before requiring matching between frames to recover camera motion. This feature extraction and matching process is often badly conditioned, noisy and not robust, and therefore it must rely on higher level robust estimation techniques and on filtering.

Direct approaches (*image-based*), however, do not rely on this feature extraction and matching process. The camera motion is directly estimated by minimising a non-linear intensity error between images, via a parametric warping function. In this way, the matching and the motion estimation are performed simultaneously at each step of the optimisation. Classically direct approaches have

focused on region-of-interest tracking whether they are modelled by affine [9], planar [10,11,12], or multiple-plane tracking [13,14]. In [15] direct approaches were generalized to use the full-image densely and track 6 DOF pose using stereo cameras whilst mapping the environment through dense stereo matching.

In general, registration is performed only between close frames (small displacements), since dense registration tasks are particularly sensible to the local convexity of the cost error function. The error function convergence depends on a number of parameters including: the inherent noise in the photometric and geometric sensor measurements, the resolution (sampling at different scales), the scene configuration (symmetry) and the scene stationarity (illumination changes or moving objects). Even though a mathematical condition for the convergence cannot be established, some effort was done in estimating convergence envelops for teaching and repeating techniques [16] [17].

We are interested in applying direct registration in larger displacements which is useful for re-location tasks, where the current trajectory may not be "close enough" to the trajectory where the model was learned and/or because the conditions of observation have changed: lighting, occlusions and dynamic objects. This problem also occurs in large scale scenes due to storage capability and complexity of configurations. In these cases RGB registration techniques have their performance challenged, since convergence is likely to happen only for small displacements (for instance see fig. 1). This work addresses a contribution in this direction by considering the information gathered from ICP point-to-plane [18] and photometric error direct cost functions [19], not only for improving ranking conditioning as in [20] and [21], but for taking into consideration the convexity of both terms to achieve a larger convergence domain and smaller number iterations.

The main contribution of this work is an adaptive RGB-D error cost function that has a larger convergence domain and a faster convergence in both simulated and real data. This formulation employs the relative condition number metric to update the weighting of the RGB and depth costs. We show that this significantly improves the convergence stability and the speed of convergence. A set of strategies are also presented to further increase the robustness of the system. First, we discuss a regularization of the geometry in planar patches that reduces the spurious noise (specially at non-textured regions) and that generates a confidence index for each pixel; Second, we present a coherent pixel selection from saliency that ensures good observation properties of each DOF for both RGB and ICP registration tasks.

The remainder is organized as follows. First, we review recent related works in Section 1.1. Next, we introduce the basic classical method of dense registration and our adapted formulation in Sections 2.1 and 2.2. Section 2.3 describes further improvements in accuracy by a regularization of the depth information and the extension of a saliency concept for computational efficiency. Lastly, we present experimental results in Section 3 for indoor (simulated and real) and outdoor contexts, and to conclude the paper in Section 4.

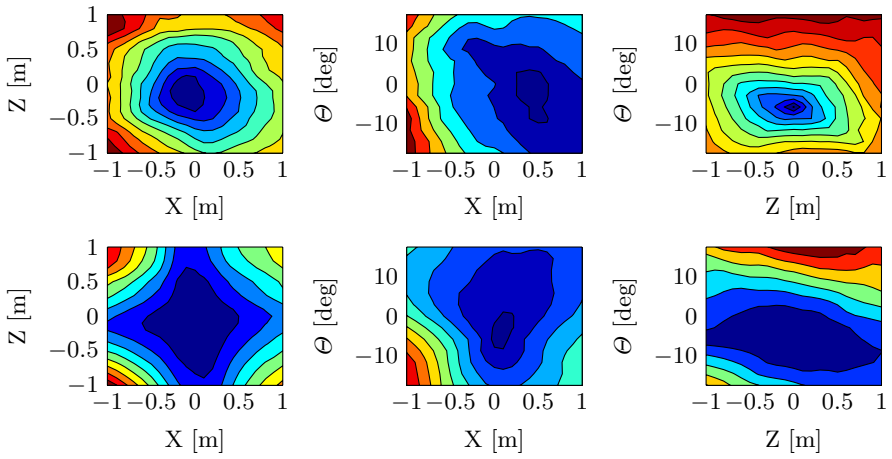

Fig. 1: Intensity RGB level curves (first row) and ICP ploint-to-plane (second row) for a typical corridor frame at the Sponza Atrium model. The costs are evaluated in the simplified case of 3DOF (one rotation and two translations) and the corresponding level curves are from the surfaces $C(\mathbf{x})$ with $\mathbf{x} = [x\ 0\ z\ \mathbf{0}_{(1\times3)}]^T$ (left column), $\mathbf{x} = [x\ 0\ 0\ \theta\ 0\ 0]^T$ (middle) and $\mathbf{x} = [0\ 0\ z\ \theta\ 0\ 0]^T$ (right column). The ICP point-to-plane cost is flatter near the solution.

1.1 Main Related Works

Large image displacement is an active area of research in the optical flow community [1] [22] [2]. Variational optimization methods are typically applied to constrain the flow estimation (each pixel has two degrees of freedom) in a dense framework. In addition, [1] jointly consider features (i.e. taking advantage of the invariance and stability of SIFT and SURF to scene changes) in cases of large motions. These approaches are not suitable in our context since we aim to keep the direct estimation concept (no matching stage).

This work is mainly related to direct RGB-D motion estimation techniques, being close to [23] and [20]. An important issue raised in these previous works is the scaling of the geometric and photometric cost terms for ensuring nice convergence properties. The interesting work of [23] applies a smooth steep function to weight the influence of the RGB and ICP terms during optimization. Although sharing a similar framework and initial conclusions (which were founded independently from their work), we propose an additional equivalent activation function based on the conditioning of the error terms. This formulation is more stable and capable of dealing with cross-peak instabilities. The work of [20] adopts λ_D to scale the photometric error (in pixels) to the geometric error (in meters) by taking $\lambda_D = median(\mathcal{I}^*)/median(\mathcal{D}^*)$. This metric ensures better ranking conditions (e.g. in cases of non-textured regions) with similar convergence rate, but fails to handle basic cases of bimodal pixel/depth intensities and the convergence properties of both costs.

Finally, our regularization method, which is an extension of [24] [25], is directly related to [26] which perform a region growing using simultaneously intensity and geometric contours. The regularization is also particularly useful in compact mapping techniques (even if not being explicitly treated in this work). Compact mapping deals with the problem of representing the world without performing an explicit 3D reconstruction of the environment in a single global reference frame [27]. This allows to create local sub-maps or to store raw (unmodified) local sensor data in the representation whilst maintaining a topological framework at large-scale that is accurate enough to ensure the connectivity between locally precise frames.

1.2 Notation and Preliminaries

A frame $\mathcal{F} = \{\mathcal{I}, \mathcal{D}\}$ is composed of an image $\mathcal{I} \in [0,1]^{m \times n}$ as pixel intensities and $\mathcal{D} \in \mathbb{R}_+^{m \times n}$ as the depth information. The mapping between the image pixel coordinates $\mathbf{p} \in \mathbb{P}^2$ and depth to 3D cartesian coordinates is given by the sensor projection $g : \mathbb{P}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}^3$. The sensor projection model of interest is the perspective and spherical model (the images are projected in the unit sphere $\mathbb{S}^2$). Point coordinates correspondences between frames are given by the warping function $w : \mathbb{P}^2 \times \mathbb{R}_+ \times \mathbb{SE}(3) \mapsto \mathbb{P}^2$, under observability conditions at different viewpoints. Denoting $\mathbf{K}$ the intrinsic sensor model and $\mathbf{q}_S \in \mathbb{S}^2$ being the unitary vector, the corresponding warping functions are given by:

$$\begin{aligned} &\bullet \text{ Perspective: } w(\mathbf{p},\mathcal{D}(\mathbf{p}),\mathbf{T}) = \frac{\mathcal{D}(\mathbf{p})\mathbf{K}\mathbf{R}\mathbf{K}^{-1}\mathbf{p}+\mathbf{K}\mathbf{t}}{[\mathcal{D}(\mathbf{p})\mathbf{K}\mathbf{R}\mathbf{K}^{-1}\mathbf{p}+\mathbf{K}\mathbf{t}]_3} \\ &\bullet \text{ Spherical: } w(\mathbf{p},\mathcal{D}(\mathbf{p}),\mathbf{T}) = \mathbf{q}_S^{-1}\left(\frac{\mathcal{D}(\mathbf{p})\mathbf{R}\mathbf{q}_S(\mathbf{p})+\mathbf{t}}{||\mathcal{D}(\mathbf{p})\mathbf{R}\mathbf{q}_S(\mathbf{p})+\mathbf{t}||}\right) \end{aligned} \tag{1}$$

where $\mathbf{q}_S^{-1}(\bullet)$ is the inverse unit sphere mapping to cartesian coordinates and the operator $[\bullet]_i$ selects the ith coordinate value. The pose $\mathbf{T}(\mathbf{x}) \in \mathbb{SE}(3)$ linking two frames (reference and target frame) is defined by the exponential map with six degrees of freedom (DOF) $\mathbf{x} \in \mathbb{R}^6$ (please see Appendix A for details). For notation convenience, in the rest of the paper $w(\mathbf{p},\mathcal{D}(\mathbf{p}),\mathbf{T}) := w(\mathbf{p},\mathbf{T})$. The normal vector of the surface s in the depth map $\mathcal{D}$, $s : \mathbb{R}^3 \mapsto \mathbb{R}$; $r - \mathcal{D}(\mathbf{p}) = 0$ is given by the gradient $\mathbf{n} = \nabla s(\mathbf{q})$, orthogonal to its tangent plane $\mathcal{P}(\mathbf{n}, d)$: $\mathbf{n}^T\mathbf{q} + d$ with $d = -\mathbf{n}^T\mathbf{q}_0, \forall \mathbf{q}_0 \in \mathcal{P}$.

2 Proposed Approach

Our adaptive RGB-D registration approach is based on classical direct VO [19] and ICP point-to-plane [18] strategies. In fact, the intensity and depth data error terms display different convergence properties for small and large motions. We aim to explore these complementary aspects, in terms of convergence, by using a modified cost function, where the geometric term prevails in the first coarse iterations, while the intensity data term dominates in the finer increments.

Next, we present additional aspects to improve the quality and robustness of this approach. They are particularly pertinent when performing localization to

previously acquired frames (e.g. when locating a target frame to a local submap). At first, the frame depth estimates are refined by taking into account the geometric and photometric continuity of the scene (using superpixels). This is done by segmenting the scene in planar patches, which improves the depth accuracy considerably whilst allowing better normal surface estimation (specially in noisy measurements from stereo). The advantages are twofold: (i) the regularization improves and reduces the spurious noise, specially at non-textured regions; (ii) the generation of a confidence index that can be used further for pixel selection in an extended saliency concept.

2.1 Hybrid RGB-D Cost Function

The pose $\hat{\mathbf{T}}\mathbf{T}(\mathbf{x})$ between a reference and a target frame is performed iteratively from a linearised convex cost function of the following photometric and geometric errors

$$e_I(\mathbf{p},\mathbf{x}) = \mathcal{I}(w(\mathbf{p},\hat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) - \mathcal{I}^*(\mathbf{p}) \tag{2}$$

$$e_D(\mathbf{p},\mathbf{x}) = \lambda_D(\hat{\mathbf{R}}\mathbf{R}(\mathbf{x})\mathbf{n}^*(\mathbf{p}))^T(g(w(\mathbf{p},\hat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) - \hat{\mathbf{T}}\mathbf{T}(\mathbf{x})g^*(\mathbf{p})) \tag{3}$$

Where $\hat{\mathbf{T}}$ is the initial pose guess; $\mathbf{n}^*$ the normal surface vector calculated at the reference frame; $g(\bullet)$ is the inverse projection model; and λ_D is a tuning parameter for scaling the error terms. The intensity only (RGB) registration method is equivalent to consider $\lambda_D = 0$. Eq. (2) is a classical optical flow constraint equation (OFCE) term (within the hypothesis of Lambertian surfaces) and (3) is equivalent to a flow point-to-plane ICP, both assuming predominant static surfaces. To ensure these assumptions, robust M-estimators (denoted as $\rho(\bullet)$) are applied for mitigating outliers influence [28]. This allows to reduce the effects of self occlusions, moving objects, illumination and interpolations errors in the direct estimation.

The classic RGB-D registration consists of using jointly (2) and (3) as

$$C(\mathbf{x}) = \sum_{\mathbf{p}} \rho_I(e_I(\mathbf{p},\mathbf{x})) + \sum_{\mathbf{p}} \rho_D(e_D(\mathbf{p},\mathbf{x})) \tag{4}$$

Choosing a large λ_D ($\lambda_D >> 1$) in (4) is equivalent to the direct ICP method, while $\lambda_D \approx 0$ corresponds to a classical dense VO. To increase the convergence rate, the optimization procedure is done considering multi-resolution Gaussian pyramidal images [19]. The optimization begins in the smallest resolution (pyramid at level n) to the bigger resolution (level 1). The corresponding Jacobians and framework is resumed in the Appendix A.

2.2 Adaptive Formulation

As stated previously in section 2.1, a main concern with direct methods is about their convergence, since only local properties are settled from eq. (2), (3) and (4). We observed in both simulated and real sequences that the intensity and

geometric terms have distinct convergence properties. While the convexity analysis of the cost terms cannot be established in general, the intensity RGB term has slower convergence (flatter) than the ICP point-to-plane cost, but its locally more precise when near the solution. This agrees with the findings of [23] in face tracking tasks. For illustration, we present typical shapes of the RGB and ICP cost terms (4) for 3DOF (two translations and one rotation) in fig. 1 of a frame in the the Sponza Atrium model dataset. The geometric error component (second row) is more discriminant than the intensity cost (first row) when further from the solution. Conversely, (as can be noticed in fig. 1) ICP is less discriminant in the vicinity of the solution, meaning that the ICP point-to-plane is flatter than the RGB term for small interframe displacements. Besides, due to the scene symmetry along the Z axis (corridor-canyon like environment), the convergence rate might be slow if the task is restricted to this DOF (see fig. 1 bottom right level plot).

Based on these observations, we propose a modified cost function where the geometric term prevails in the first coarse iterations, while the intensity data term dominates in the finer increments (in the neighbourhood of the solution). A natural candidate activation function $\mu(\mathbf{x})$ in this context is the smoothed step

$$\mu(\mathbf{x}) = k_1 / \left(1 + \exp(-k_2(||\mathbf{x}|| - c))\right) \tag{5}$$

depending on the size of the pose increments $\mathbf{x}$ along the minimization of (4). This is similar to the solution adopted in [23]. Note that selecting a high value to k_2 and c near the numeric iteration limit is equivalent to perform a sequential independent ICP and intensity RGB tasks (in cascade). Therefore, the activation (5) is particularly sensitive to the tuning parameters k_2 and c, which can induce oscillations (cross-peak) by transforming the original cost (4) in a non convex function. To address these issues, a second strategy is to analyse the costs relative

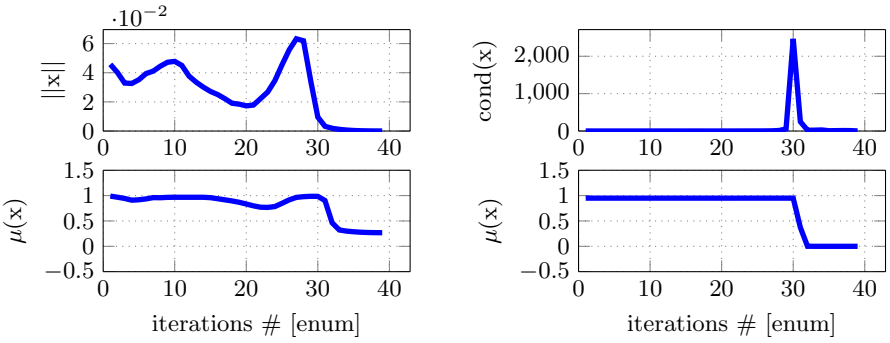

Fig. 2: Activation adaptive function $\mu(\mathbf{x})$ while performing registration in the KITTI outdoor dataset. The left column corresponds to (5) and the right to (6) with the tuning parameters of Table 1. The conditioning criteria (6) is easily detectable. Note that the norm of the pose increments using (5) are not monotonically decreasing (top left).

behaviour along the optimization steps. The idea is that the relative conditioning number detects when the algorithm is in the vicinity of the solution (i.e. where the ICP cost is less discriminant). An equivalent adaptive function is then:

$$\mu(\mathbf{x}) = k_1 \mathbf{1}(cond_{\mathbf{x}}(C_I(\mathbf{x}))/cond_{\mathbf{x}}(C_D(\mathbf{x}))) \tag{6}$$

where the indicator function $\mathbf{1}(\bullet)$ is zero when $cond_{\mathbf{x}}(C_I(\mathbf{x}))/cond_{\mathbf{x}}(C_D(\mathbf{x})) >> 1$ and one otherwise and

$$cond_{\mathbf{x}}(C(\mathbf{x})) = \left| \frac{C(\mathbf{x}_0 \circ \mathbf{x}) - C(\mathbf{x}_0)}{C(\mathbf{x}_0)} \right| / \frac{||\mathbf{x}||}{||\mathbf{x}_0||} \tag{7}$$

being an estimate of the relative condition number of the RGB (C_I) and ICP (C_D) cost functions, with $\mathbf{x}_0 = se3^{-1}(\log(\hat{\mathbf{T}}))$ and $\circ$ is the additive Lie algebra action. We show in fig. 2 typical curves using the KITTI VO/SLAM dataset [29] for both adaptive metrics – eq. (5) is presented in the first column and (6) in the second. The parameters of each activation are detailed in Table 1. This activation proved to detect correctly the sensitivity of the costs, whilst being of easy tuning. The respective hybrid modified cost function is designed as a joint adaptive RGB-ICP cost

$$\tilde{C}(\mathbf{x}) = (1 - \mu(\mathbf{x})) \sum_{\mathbf{p}} \rho_I(e_I(\mathbf{p}, \mathbf{x})) + \mu(\mathbf{x}) \sum_{\mathbf{p}} \rho_D(e_D(\mathbf{p}, \mathbf{x})) \tag{8}$$

Where the respective Jacobians are linear combinations of the Jacobians from the original formulation: $\tilde{\mathbf{J}} = [\sqrt{(1 - \mu(\mathbf{x}))}\mathbf{J}^I \quad \sqrt{\mu(\mathbf{x})}\mathbf{J}^D]^T$. Finally, we adopted the Huber robust function for ρ_I, ρ_D for ensuring convexity properties when further from the solution [28]. To avoid outliers influence, the robust function is switched to Tukey when in the vicinity of the minimum (i.e. when the conditioning in (6) is large). The respective Jacobians and details about the optimization are given in the Appendix A. As it will be shown later, this formulation proved to have significant advantages in extensive tests performed in simulated and in real indoor and outdoor sequences.

2.3 Regularization in Planar Patches and Pixel Selection

Before estimating inter-frame poses, we can perform a regularization based on the assumption that the scene contains piecewise smooth surfaces. The frame depth map is then represented as a set of planes of variable size, where non-planar surfaces are approximated by a set of small planar patches. This assumption is applicable for any environment: structured and non-structured as long as as the planar approximation error is smaller than the measurement error. This can be easily achieved by selecting the suitable parameters according to the sensor noise model (regardless the type of scene).

The plane segmentation is performed by region growing, starting from a set of seeds distributed around the image. The conditions are: (i) in order to add a pixel i into a group of neighbouring pixels representing a surface s, we

Fig. 3: Spherical intensity image (left), point cloud (middle) and segmented patches (right) of rendered indoor scene using the superpixels constraints. For display purposes, plane colors are the average color of each patch. The segmentation preserves edges as the calibration target and whilst reducing the noise in frontier stitched regions.

verify if their normal vector have the similar direction; and (ii) that the 3D point corresponding to the pixel $\mathbf{p}$ lies approximately on the plane defined by s (orthogonal projection error).

A last stage is carried out to merge the contiguous planar patches that lie approximately onto the same 3D plane. For that we exploit the color continuity of the images, where changes in color or texture often delimit different objects and different planar surfaces. This is used to guide the planar segmentation of the scene and to reduce the smoothing/aliasing effect of objects. The color image is segmented in superpixels with similar photometric properties (color and texture) using the single linear iterative clustering (SLIC) algorithm [30], which encodes nice properties as strong adherence to boundaries and compactness. Afterwards, the mean patch $(d_s, \mathbf{n}_s)$ related to each superpixel region is extracted considering all the patches that are mostly englobed by that superpixel. The assumption about the planar continuity of the scene will help to reduce the uncertainty since new information is provided. Thus, the uncertainty of a depth measurement (pixel in the depth image) belonging to a plane is modelled as:

$$\Sigma_{\mathcal{D}}(\mathbf{p}) = \frac{r_\sigma(\mathbf{p})}{|\mathbf{n}_s^T \mathbf{q}_S(\mathbf{p})|} \tag{9}$$

where r_σ is the ratio between the smallest and the bigger eigenvalue of the covariance of the patch (the smallest eigenvalue from the singular value decomposition of the covariance matrix of the 3D points).

Observability and Information Selection In this section we derive the concepts for exploiting the certainty index gathered from the regularization along with the pixel selection. Since only a subset of the image information is useful for pose computation, e.g. non textured regions has no influence in RGB registration cost term – no photometric gradient. The problem is then to select points that best constraints the cost functions and discard the ones who does not This formulation is equivalent to ensure that the Fisher Information Matrix

(FIM) $\mathbf{J}^T\mathbf{J}$ is well conditioned. This is performed by simply analysing the magnitude of the analytical Jacobians columns (see Appendix A for details). This is the underlying idea presented in the works of [18] for the ICP cost and in [31] for the RGB cost (2). Hence, the update row ranking considering the geometric confidence is done under the following modified Jacobian

$$\bar{\mathbf{J}} = \frac{1}{\Sigma_D}\begin{bmatrix} (||\mathbf{J_1}^I|| + ||\mathbf{J_1}^D||) & ||\mathbf{J_2}^I|| + ||\mathbf{J_2}^D||) & ||\mathbf{J_3}^I|| + ||\mathbf{J_3}^D||) \\ ||\mathbf{J_4}^I|| + ||\mathbf{J_4}^D||) & ||\mathbf{J_5}^I|| + ||\mathbf{J_5}^D||) & ||\mathbf{J_6}^I|| + ||\mathbf{J_6}^D||) \end{bmatrix} \quad (10)$$

with Σ_D the plane uncertainty index (9). The advantages are twofold: (i) to exclude the points that does not contributes directly to the estimation (the valid information can be masked from the spurious noise in the redundant not useful information); (ii) computation efficiency whilst keeping the precision.

3 Experiments and Results

We evaluate the technique in sequences of indoor and outdoor images using perspective and spherical RGB-D sensors. We consider the average of the rotation relative error (RRE), translation relative error (TRE) and number of iterations to convergence as quantitative metrics. In cases of lack of ground truth (e.g. in the real indoor spherical sequence), a qualitative analysis is done for the intensity and depth errors. Unless specified, the term Adaptive RGB-D corresponds to the cost (8) using (6).

Implementation Aspects The iterative pose estimation algorithm (see Appendix A) is said to have converged (either to a global or to a local minimum) when the norm of pose increments $\mathbf{x}$ are bellow a fixed threshold in successive iterations (10^{-5} for the rotation and 10^{-3} for translation). The parameters employed (λ_D and activation function (5)) are described in Table 1. We used the same parameters in all the next experiments.

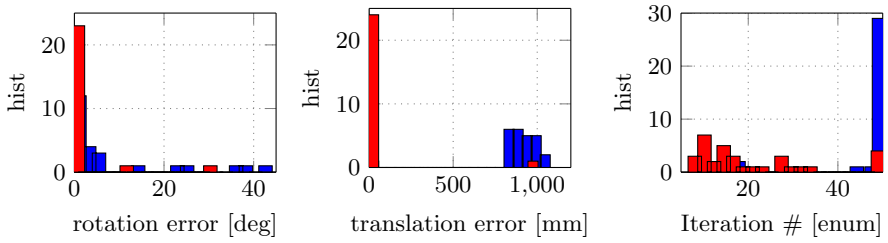

Fig. 4: Rotation, translation errors and number of iterations for a fixed resolution for the simulated testbed dataset with gap of 10 frames. RGB-D dense registration errors in rotation (left in degrees), translation (center in millimetres) and number of iterations to converge (right figure). The registration considering the classical RGB-D is presented in blue, while the adaptive formulation using the conditioning (6) is in red. The precision and convergence rate are substantially improved when exploiting the activation factor (in red).

Table 1: Parameters in the activation functions.

	Parameters	Typical Range
Meth. 1 [20]	$\lambda_D = med(\mathcal{I})/med(\mathcal{D})$, $\mu = 0.5$	$\lambda_D \in [5, 50]$
Adapt. 1 (5)	$\lambda_D = 1$, $k_1 = 1 - 10^{-5}$, $k_2 = 100$, $c = 0.001$	$\mu \in [0, k_1]$
Adapt. 2 (6)	$\lambda_D = 1$, $k_1 = 1 - 10^{-5}$	$\mu \in \{0, k_1\}$

– **Spherical Simulated Sequence:**
At first, we evaluate our approach in controlled conditions using 500 RGB-D spherical synthesized images from the Sponza Atrium model. We start in a fixed resolution for evidencing the differences between the classic RGB-D and the adaptive formulation. The maximum number of iterations is 50. To emulate different motion speeds, only a sub-set of the frames is picked up (gaps of 5, 10 and 15 frames) – fig. 4 shows the pose errors and the number of iterations for a gap of 10 spheres. The registration results are synthesized in Table 2. The distances between frames are in average of 0.15 [m] and of 4 degrees in rotation. The convergence was achieved even in cases considering translations and rotations of around 2.5[m] and 60[deg]. The convergence failed in less than 10% of trials in the furthest experiment (gap of 15). These cases happened when the reference scene was almost completely occluded in the target scene (e.g. corridor 90 degrees turns) and are expected to happen since the direct method's hypothesis of overlapping is not fulfilled.

Table 2: Quantitative simulated spherical indoor sequence in a fixed resolution: average RRE[deg]/RTE[mm]/Iterations.

	Gap = 5	*Gap = 10*	*Gap = 15*
Meth. 1 [20]	3.67/423/47.3	7.80/1104/48.4	11.7/1520/48
Adapt. 1 (5)	0.68/96.4/31.2	1.11/466/32.9	2.17/833/34.8
Adapt. 2 (6)	**0.03/88.6/31**	**0.04/182/26.5**	**0.05/523/20.7**

– **Spherical Indoor and Outdoor Real Sequences:**
The spherical indoor and outdoor frames are acquired using two designed spherical RGB-D sensors [32] [31]. The indoor images were captured in the hall and offices of the Inria building using a set of eight Asus sensors. A more qualitative analysis is done due to the lack of ground truth. With a separation of five frames, the method did not converge in only 9% of the trials. They correspond mostly to the cases where the maximum number of iterations was reached – in this case 150 iterations as in fig. 5. The adaptive solution had a better performance with a much smaller number of iterations. Note that we apply the same experiment to the RGB only cost function (i.e. eq. (4) with $\lambda_D = 0$), but it either reached a local minima or did not converge for most of the frames (black trajectory in fig. 5 left). The same experiment was

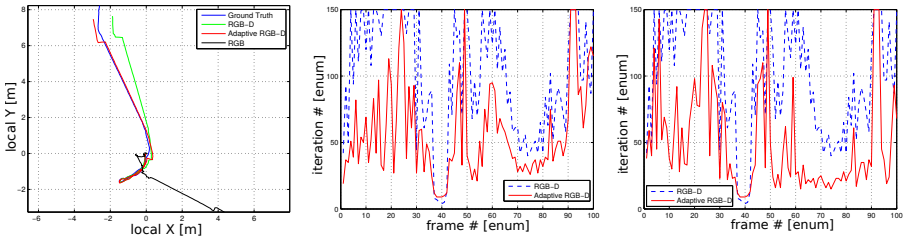

Fig. 5: Trajectories (left) using RGB, RGB-D and adaptive RGB-D over a sequence of 500 frames. A gap of five frames were used to compute each task. The ground truth was obtained using the RGB with step of one frame. The number of iterations to convergence with the adaptive RGB-D (red curves in the center) is significantly smaller than when considering the classic method (in blue). The number of iterations is reduced when taking into account the regularization+saliency stage.

performed with the regularized+saliency criteria selecting the most informative points. Although the convergence success rate remained invariant, when compared to the adaptive one, it gives more stability and efficiency to the optimization since a reduced number of iterations are performed (see right fig. 5). Finally, we depict two experiments in the presence of large motions and dynamics objects (see fig. 6). The data was acquired in the two different regions of a building with many occlusions, large rotations, and with dynamic objects. The same conclusions were obtained from the outdoor data, acquired in an urban/semi-urban area using a spherical stereo system [31]. A qualitative view of the respective intensity and geometric errors during a registration task can be seen in fig. 7.

– **KITTI Outdoor Perspective Sequence**
We also provide results for the perspective outdoor sequence of the KITTI Visual Odometry/SLAM benchmark. It is a challenging dataset since the scene is mainly semi-structured (roads in an urban area) and with a travel speed up to 60km/h. We observed that in the outdoor scenarios the overlapping regions are much sparser because only the road plane is the persistent overlapping surface. The respective error metrics are displayed in Table 3 for a fixed resolution.

Table 3: Quantitative KITTI outdoor sequence in fixed resolution: average RRE[deg]/RTE[mm]/Iterations.

	Gap = 1	*Gap = 2*	*Gap = 3*
Meth. 1 [20]	0.51/219/45.6	1.83/1071/49	2.75/1846/50
Adapt. 1 (5)	0.27/120/36.5	1.12/557/45	2.34/1101/**46.7**
Adapt. 2 (6)	**0.08/35.1/33.5**	**0.42/192/41.7**	**1.79/825**/47

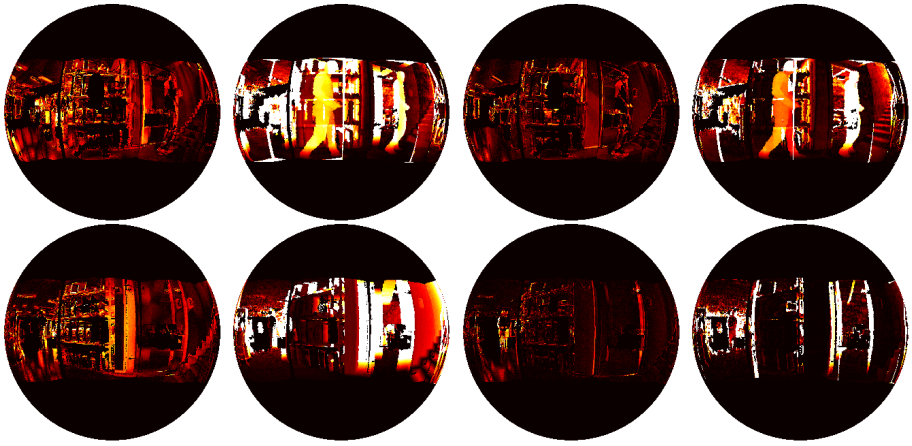

Fig. 6: Intensity and geometry errors between indoor frames with a gap of fifteen frames for two different scenes. Each row is composed of two pairs of errors: the classic RGB-D and the adaptive formulation in the form $(\mathbf{e_I}(\mathbf{x}), \mathbf{e_D}(\mathbf{x}))$.

Lastly, we combined the adaptive formulation with a multi-resolution Gaussian pyramid of four levels (the higher the level, the smaller the image resolution is) to assets the efficiency of the approach in this context (see Table 4). The maximum number of iterations was of 50 at each pyramid level. To account the different computational cost of one iteration between the levels, we define the total number of iterations as $\sum_{i=1}^{4} l_i(2^{4-i})^2$ (with l_i the number of iterations at level i). The adaptive formulation is still more efficient and precise, although the discrepancy between the methods is reduced.

Table 4: Quantitative KITTI outdoor sequence results using multi-resolution (pyramid of four levels): average RRE[deg]/RTE[mm]/Iterations.

	Gap = 1	*Gap = 2*	*Gap = 3*
Meth. 1 [20]	0.08/23.1/**447**	0.78/268/980	3.68/1059/1872
Adapt. 1 (5)	**0.06**/**16.5**/704	**0.19**/81.4/**856**	**0.83**/251/**1078**
Adapt. 2 (6)	**0.06**/**16.4**/1102	0.37/**47.5**/1269	1.05/**238**/1473

4 Conclusions

In this paper, we have presented an efficient RGB-D registration approach in the context of large inter-frame displacements. The technique exploits adaptively the photometric and geometric error terms based on their convergence

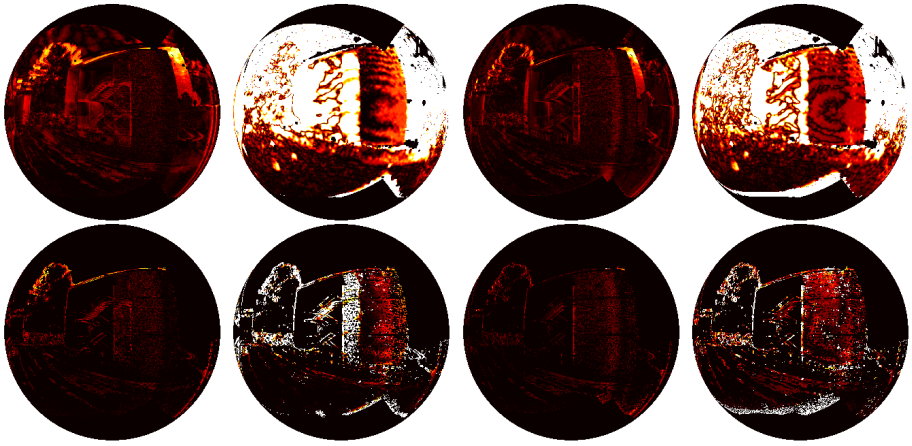

Fig. 7: Intensity and depth errors between outdoor frames with a gap of fifteen frames. The first two columns are of a classic RGB-D and the last two columns correspond to the adaptive approach. Each row is composed of the final errors: intensity $(\mathbf{e_I}(\mathbf{x}), \mathbf{e_D}(\mathbf{x}))$. The adaptive RGB-D (last columns) has smaller geometric and photometric errors (bigger errors are encoded with lighter colors). The regularization + saliency had particularly improved the resulting pose computation for the outdoor case (second row).

characteristics. Additional aspects as a two step regularization and an extended pixel saliency selection improved the quality and robustness of this approach. Despite its simplicity, this technique was capable to deal with large rotations, occlusions and moving objects in real indoor and outdoor scenarios.

Future directions includes: (i) the formal characterization of the convergence domain for different symmetries and noise statistics for both intensity and geometry data terms; and (ii) finding convex (quasi-convex) dual formulations for adding more stable dense/semi-dense features as planes, edgelets and image moments in both intensity and geometric terms.

Acknowledgements. The authors thank Josh Picard and Paolo Salaris for the discussions/proof reading of the manuscript, and the reviewers for their thoughtful comments. This work was funded by CNPq of Brazil under contract number 216026/2013-0.

Appendix A: Error Jacobians and Optimization

The pose $\mathbf{T}(\mathbf{x}) \in \mathbb{SE}(3)$ is parametrized as function of angular and linear velocities $\mathbf{x} = (\boldsymbol{\upsilon}\delta t, \boldsymbol{\omega}\delta t) \in \mathbb{R}^6$ and the optimization will be related to this twist parametrization. The pose is related to the twist velocities by the exponential

mapping $\mathbf{T}(\mathbf{x}) = \exp(se3(\mathbf{x}))$, with

$$se3(\mathbf{x}) = \begin{bmatrix} \mathbf{S}(\boldsymbol{\omega})\delta t & \boldsymbol{\upsilon}\delta t \\ \mathbf{0}_{(1\times 3)} & 0 \end{bmatrix} \in \mathfrak{se}(3) \tag{11}$$

which is the Lie algebra of $\mathbb{SE}(3)$ at the identity element, $\mathbf{S}(\mathbf{z})$ represents the skew symmetric matrix associated to vector $\mathbf{z}$ and $\delta t = 1$.

The respective Jacobians will be derived following this parametrization. We ask the reader to see [19] for details about the photometric Jacobian $\mathbf{J}^I$. Next, for the geometric point-to-plane direct Jacobian $\mathbf{J}^D \in \mathbb{R}^{1\times 6}$, we denote the 3D point error $\zeta(\mathbf{x})$:

$$\begin{aligned} \zeta(\mathbf{x}) &= -\hat{\mathbf{T}}\mathbf{T}(\mathbf{x}) \begin{bmatrix} g^*(\mathbf{p})) \\ 1 \end{bmatrix} + g(w(\mathbf{p}, \hat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) \\ &= -\hat{\mathbf{R}}\mathbf{R}(\mathbf{x})g^*(\mathbf{p}) - \hat{\mathbf{R}}\mathbf{t}(\mathbf{x}) - \hat{\mathbf{t}} + g(w(\mathbf{p}, \hat{\mathbf{T}}\mathbf{T}(\mathbf{x}))) \end{aligned} \tag{12}$$

From eqs. (3), (12) and the product rule:

$$\mathbf{J}^D(\mathbf{0}) = \lambda_D \mathbf{n}^{*T} \left(\frac{\partial(\mathbf{R}(\mathbf{x})^T \hat{\mathbf{R}}^T \zeta(\mathbf{z}))}{\partial \mathbf{x}} \bigg|_{\mathbf{z}=\mathbf{x}} + \mathbf{R}(\mathbf{x})^T \hat{\mathbf{R}}^T \frac{\partial(\zeta(\mathbf{x}))}{\partial \mathbf{x}} \right) \bigg|_{\mathbf{x}=\mathbf{0}} \tag{13}$$

For clarity, the first term in eq. (13) is $\mathbf{J_{d1}}$ and we decompose the second term in two Jacobians $\mathbf{J_{d2}}$ and $\mathbf{J_{d3}}$, such as $\mathbf{J}^D(\mathbf{0}) = \lambda \mathbf{n}^{*T}(\mathbf{J_{d1}}(\mathbf{0}) + \mathbf{J_{d2}}(\mathbf{0}) + \mathbf{J_{d3}}(\mathbf{0}))$. From $\frac{\partial(\mathbf{R}(\mathbf{x})\zeta)}{\partial \mathbf{x}} = \frac{\partial(\mathbf{R}(\mathbf{x})\zeta)}{\partial \mathbf{R}(\mathbf{x})} \frac{\partial \mathbf{R}(\mathbf{x})}{\partial \mathbf{x}}$ the first term is

$$\mathbf{J_{d1}}(\mathbf{0}) = \begin{bmatrix} \mathbf{0}_{3\times 3} & \mathbf{S}(\hat{\mathbf{R}}^{\mathbf{T}}\zeta(\mathbf{0})) \end{bmatrix} \tag{14}$$

The second term is decomposed in two Jacobians

$$\mathbf{J_{d2}}(\mathbf{0}) = \begin{bmatrix} -\mathbf{I}_{3\times 3} & \mathbf{S}(g^*(\mathbf{p})) \end{bmatrix} \tag{15}$$

And finally the last Jacobian is the one corresponding to $\frac{\partial(g(w(\mathbf{p}, \hat{\mathbf{T}}\mathbf{T}(\mathbf{x})))}{\partial \mathbf{x}}$. This derivative can be seen as an extended version of the image photometric gradient $\mathbf{J^I}$, for each component of $g(w(\mathbf{p}, \hat{\mathbf{T}}\mathbf{T}(\mathbf{x}))$. Then

$$\mathbf{J_{d3}}(\mathbf{0}) = \begin{bmatrix} \mathbf{J_g}\big|^T_{[g(\mathbf{p}_w))]_1} & \mathbf{J_g}\big|^T_{[g(\mathbf{p}_w))]_2} & \mathbf{J_g}\big|^T_{[g(\mathbf{p}_w))]_3} \end{bmatrix}^T \mathbf{J_w} \mathbf{J_T} \tag{16}$$

And $\mathbf{J_g}\big|_{[g(\mathbf{p}_w))]_i}$ is the image gradient (as in the photometric term) of an image produced with the ith-coordinate of $g(w(\mathbf{p}, \hat{\mathbf{T}}\mathbf{T}(\mathbf{0})))$. Note that this Jacobian is small for points belonging to planar surfaces. Therefore, $\mathbf{J_{d3}}$ is neglected since only a fraction of the scene is on geometric discontinuities and since these points have higher sensitivity to depth error estimates and self-occlusions effects. Finally, we use the ESM formulation [19] for defining the optimization step for the RGB cost, while a Gauss-Newton step is employed for the geometric Jacobian. The reader is asked to see [33] for more details on the different optimization available techniques.

References

1. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE PAMI **33** (2011)
2. Braux-Zin, J., Dupont, R., Bartoli, A.: A general dense image matching framework combining direct and feature-based costs. In: IEEE ICCV. (2013)
3. Howard, A.: Real-time stereo visual odometry for autonomous ground vehicles. In: IEEE IROS. (2008)
4. Davison, A., Murray, D.: Simultaneous localization and map-building using active vision. IEEE TPAMI **24** (2002)
5. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: IEEE CVPR. (2004)
6. Kitt, B., Geiger, A., Lategahn, H.: Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In: IEEE IV. (2010)
7. Harris, C., Stephens, M.: A combined corner and edge detector. In: 4th Alvey Vision Conference. (1988)
8. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004)
9. Hager, G., Belhumeur, P.: Efficient region tracking with parametric models of geometry and illumination. IEEE TPAMI **20** (1998)
10. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI. (1981)
11. Irani, M., Anandan, P.: Robust multi-sensor image alignment. In: ICCV. (1998)
12. Baker, S., Matthews, I.: Equivalence and efficiency of image alignment algorithms. In: IEEE CVPR. (2001)
13. Mei, C., Benhimane, S., Malis, E., Rives, P.: Constrained multiple planar template tracking for central catadioptric cameras. In: BMVC. (2006)
14. Caron, G., Marchand, E., Mouaddib, E.: Tracking planes in omnidirectional stereovision. In: IEEE ICRA. (2011)
15. Comport, A., Malis, E., Rives, P.: Accurate quadrifocal tracking for robust 3d visual odometry. In: IEEE ICRA. (2007)
16. Churchill, W., Tong, C., Gurau, C., Posner, I., Newman, P.: Know your limits: Embedding localiser performance models in teach and repeat maps. In: IEEE ICRA. (2015)
17. Furgale, P., Barfoot, T.: Visual teach and repeat for long-range rover autonomy. JFR **27** (2010)
18. Gelfand, N., Ikemoto, L., Rusinkiewicz, S., Levoy, M.: Geometrically stable sampling for the icp algorithm. In: 3DIM. (2003)
19. Comport, A., Malis, E., Rives, P.: Real-time quadrifocal visual odometry. IJRR **29** (2010)
20. T.Tykkala, Audras, C., Comport, A.: Direct iterative closest point for real-time visual odometry. In: ICCV Workshops. (2011)
21. Kerl, C., Sturm, J., Cremers, D.: Dense visual SLAM for RGB-D cameras. In: IEEE IROS. (2013)
22. Timofte, R., Gool, L.V.: Sparse flow: Sparse matching for small to large displacement optical flow. In: IEEE WCACV. (2015)
23. Morency, L., Darrell, T.: Stereo tracking using icp and normal flow constraint. In: ICPR. (2002)
24. Martins, R., Fernandez-Moral, E., Rives, P.: Dense accurate urban mapping from spherical RGB-D images. In: IEEE IROS. (2015)

25. Gokhool, T., Martins, R., Rives, P., Despre, N.: A compact spherical RGBD keyframe-based representation. In: IEEE ICRA. (2015)
26. Weikersdorfer, D., Gossow, D., Beetz, M.: Depth-adaptative superpixels. In: IEEE ICP. (2013)
27. Fernandez-Moral, E., Mayol-Cuevas, W., Arevalo, V., Gonzalez-Jimenez, J.: Fast place recognition with plane-based maps. In: IEEE ICRA. (2013)
28. Zhang, Z.: Parameter estimation techniques: A tutorial with application to conic fitting. Technical Report 2676, Inria (1995)
29. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE CVPR. (2012)
30. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixels methods. IEEE Trans. PAMI **34** (2012)
31. Meilland, M., Comport, A., Rives, P.: Dense omnidirectional RGB-D mapping of large-scale outdoor environments for real-time localization and autonomous navigation. JFR **32** (2015)
32. Fernandez-Moral, E., Gonzalez-Jimenez, J., Rives, P., Arevalo, V.: Extrinsic calibration of a set of range cameras in 5 seconds without pattern. In: IEEE IROS. (2014)
33. Barker, S., Matthews, I.: Lucas-kanade 20 years on: a unifying framework. IJCV **56** (2006)